A fast method for high-resolution voiced/unvoiced detection and glottal closure/opening instant estimation of speech

Koutrouvelis, A.I.; Kafentzis, GP; Gaubitch, ND; Heusdens, R

DOI
10.1109/TASLP.2015.2506263

Publication date
2015

Document Version
Accepted author manuscript

Published in
IEEE - ACM Transactions on Audio, Speech, and Language Processing

A Fast Method for High-Resolution Voiced/Unvoiced Detection and Glottal Closure/Opening Instant Estimation of Speech

Andreas I. Koutrouvelis, George P. Kafentzis, Nikolay D. Gaubitch, and Richard Heusdens

***Abstract*—We propose a fast speech analysis method which simultaneously performs high-resolution voiced/unvoiced detection (VUD) and accurate estimation of glottal closure and glottal opening instants (GCIs and GOIs, respectively). The proposed algorithm exploits the structure of the glottal flow derivative in order to estimate GCIs and GOIs only in voiced speech using simple time-domain criteria. We compare our method with well-known GCI/GOI methods, namely, the dynamic programming projected phase-slope algorithm (DYPSA), the yet another GCI/GOI algorithm (YAGA) and the speech event detection using the residual excitation and a mean-based signal (SEDREAMS). Furthermore, we examine the performance of the aforementioned methods when combined with state-of-the-art VUD algorithms, namely, the robust algorithm for pitch tracking (RAPT) and the summation of residual harmonics (SRH). Experiments conducted on the APLAWD and SAM databases show that the proposed algorithm outperforms the state-of-the-art combinations of VUD and GCI/GOI algorithms with respect to almost all evaluation criteria for clean speech. Experiments on speech contaminated with several noise types (white Gaussian, babble, and car-interior) are also presented and discussed. The proposed algorithm outperforms the state-of-the-art combinations in most evaluation criteria for signal-to-noise ratio greater than 10 dB.**

***Index Terms*—Glottal closure instants (GCIs), glottal opening instants (GOIs), pitch estimation, speech analysis, voiced/unvoiced detection (VUD).**

I. Introduction

THE accurate estimation of the timing of vocal fold closure (and less often, that of vocal fold opening) during voiced speech is an important module in many speech-related technologies. In speech analysis nomenclature, these timing instants are called glottal closure instants (GCIs) and glottal opening instants (GOIs). Applications of GCI and GOI estimation are numerous, including pitch tracking [1], [2], voice source modeling [3]–[6], speech enhancement [7], closed-phase analysis and glottal flow estimation [8]–[11], speaker identification [9], [12], [13], speech dereverberation [14], speech synthesis [15], [16], speech coding [17], speech modification [18], [19] and speech transformations [20].

Several methods have been proposed for GCI estimation [2], [21]–[28], but only a few for both GCI and GOI [8], [9], [29]–[32] or GOI only estimation [33]. To the authors' knowledge, the sliding linear prediction covariance analysis [8] was the first method proposed for GCI/GOI estimation. It uses a sliding covariance analysis window that moves forward one sample at a time, and a function of the linear prediction (LP) residual to detect the closed-phase interval. In the Hilbert envelope method [29], the GCIs and GOIs are estimated using the peaks of the Hilbert envelope of the LP residual. The dynamic programming projected phase-slope algorithm (DYPSA) [25], [30] is a GCI estimation method that uses the phase slope function of the residual to extract candidate GCIs and then performs N-best dynamic programming to obtain an optimal GCI set. DYPSA also estimates GOIs by using a fixed closed-quotient interval of 0.3 s [34]. The speech event detection using the residual excitation and a mean-based signal (SEDREAMS) algorithm [31], [35] estimates GCIs and GOIs from the sharp epochs of the LP residual in fixed intervals around the zero-crossings of a mean-based signal. The latter is a smoothed, windowed version of the speech signal and the window length is a function of the mean pitch of the speech signal. The yet another GCI/GOI algorithm (YAGA) [32] follows a similar strategy to DYPSA based on the phase slope function and on N-best dynamic programming, but differs in two main ways. YAGA applies the phase slope function on the wavelet transform of the source signal in order to emphasize the discontinuities in GCIs and GOIs. Then, it finds the best candidate set of GCIs through N-best dynamic programming and, subsequently, estimates the most consistent corresponding GOIs according to their closed-quotient.

GCIs/GOIs are meaningful only in voiced speech regions and, thus, voiced/unvoiced detection (VUD) must be applied in conjunction with GCI/GOI algorithms. Several VUD algorithms have been proposed in the literature [36]–[42]. The robust algorithm for pitch tracking (RAPT) [39] and the summation of residual harmonics (SRH) [41] appear to be the state-of-the-art methods in clean and noise-contaminated speech, respectively [41]. Both algorithms are frame-based and, therefore, do not have good resolution at voiced segment boundaries. All the previously discussed GCI/GOI algorithms estimate GCIs/GOIs in the entire speech signal (in both voiced and unvoiced segments). It is worth noting that YAGA also has a "voiced-only" version which eliminates the estimated GCIs/GOIs of unvoiced regions in an additional step [32]. To the best of our knowledge, there are no previous experimental evaluations of combined VUD and GCI/GOI algorithms which can reveal possible bottlenecks that deteriorate performance in real-world applications. This kind of evaluation is performed in the current work.

We propose the glottal closure/opening instant estimation forward-backward algorithm (GEFBA), which performs simultaneous high-resolution VUD and GCI/GOI estimation. Compared to the majority of the state-of-the-art approaches that are based on the LP residual, GEFBA operates on the

source signal itself, obtained by simple LP-based inverse filtering, using simple time-domain criteria.

In Figure 1 we depict, from top to bottom, a) a voiced speech segment, b) the derivative of the electroglottograph signal (dEGG), c) the LP residual, and d) the source signal, i.e. the glottal flow derivative (GFD). The reason for using the source signal instead of the LP residual in our work is two-fold. Firstly, the GFD has a convenient structure which can be exploited to identify the voiced segments. Secondly, in voiced segments with low vocal intensity (see the time interval $[20, 60]$ ms in Figure 1) the residual does not have sharp epochs and, therefore, the identification of GCIs and GOIs is difficult. This problem becomes harder when noise is added to the speech signal. On the contrary, as Figure 1(d) demonstrates, the GFD waveform suffers less from such problems due to its clearer structure. Even in SEDREAMS, which combines a smooth signal with the LP residual, the problem remains, even though the error is bounded by the length of the interval around the zero-crossings [35]. GEFBA does not need such a refinement because it uses a smooth signal which, by definition, can give the locations of GCIs/GOIs. It is worth noting that YAGA also estimates the source signal, but it does not use the source signal itself for GCI/GOI estimation. Instead, it finds discontinuities of the source signal which may not exist in some voiced segments as shown in Figure 1. It can be observed that the dEGG has large epochs even in regions where the LP residual does not. There are cases of voiced segments at which the dEGG might not have distinguishable epochs. Specifically, in voicing offsets the vocal folds may still oscillate without getting close enough to register an epoch in the dEGG [43]. However, the GFD structure does not explicitly depend on the contact of the vocal folds. Therefore, it remains quasi-periodic enabling the correct identification of GCIs/GOIs in these cases.

GEFBA achieves high-resolution VUD for two main reasons: a) the GFD waveform can reveal the voicing offsets, as already discussed and b) it is a pitch-period-based rather than a frame-based VUD. Moreover, GEFBA, using simple time-domain criteria based on the estimated glottal parameters (GCIs, GOIs etc.), can distinguish voiced from unvoiced segments by taking advantage of the similarity of the neighbouring glottal pulses. Finally, GEFBA has low complexity due to its simple LP-based inverse filtering scheme and the simple time-domain criteria used in the joint VUD and GCI/GOI estimation.

Experiments on clean speech from the SAM [44] and APLAWD [45] databases show that GEFBA outperforms the state-of-the-art combinations of VUD and GCI/GOI algorithms with respect to most evaluation criteria. In particular, it has the highest identification ratio, a remarkably better GOI accuracy, and a much lower computational complexity. GEFBA is evaluated for three different types of additive noise: white Gaussian noise (WGN), babble noise and car-interior noise. In the presence of WGN, GEFBA outperforms the state-of-the-art combinations of VUD and GCI/GOI algorithms with respect to most evaluation criteria. For the other two types of noise, GEFBA provides robust results mostly for moderate and high signal-to-noise ratios (SNRs) (i.e., above 10 dB). The source code for GEFBA can be found online[1].

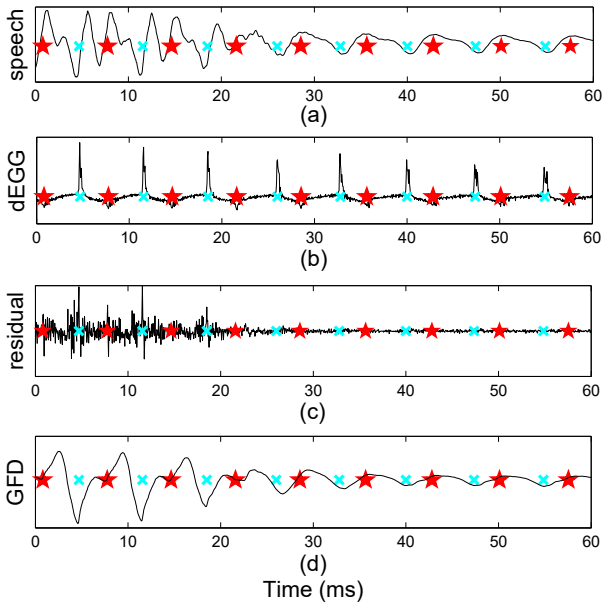


Fig. 1. A justification for using the glottal flow derivative (GFD) for the estimation of the glottal closure instants (GCIs) and the glottal opening instants (GOIs): (a) speech segment, (b) derivative of the electroglottograph (dEGG) signal, (c) linear prediction (LP) residual, (d) GFD. Stars and 'x'-marks denote the reference GOIs and GCIs, respectively, obtained from the dEGG peaks. It is clear that the LP residual is not suitable for estimating GCIs/GOIs (after 20 ms) because it does not have distinguishable epochs. On the contrary, the GFD has a smooth and clear structure, allowing a more convenient estimation of glottal instants.

The remainder of this paper is organized as follows: In Section II, the problem formulation is presented. Section III presents the GEFBA algorithm. In Section IV, the GEFBA algorithm is compared to the state-of-the-art combinations of VUD and GCI/GOI algorithms and in Section V, the results of the comparisons are discussed. Finally, we draw conclusions in Section VI.

II. Problem Formulation

A popular model for speech production is the source-filter model [46], [47]. According to this model, speech is generated as the manifestation of a source signal coming out from the vocal folds, passing through the vocal tract, and finally modified by the lip radiation. A speech signal can be classified as voiced or unvoiced depending on the state of the vocal folds (oscillating or not). In voiced speech, the vocal folds oscillate, thus producing a quasi-periodic source signal named the *glottal flow*. In unvoiced speech the vocal folds remain open and a constriction is formed in certain parts of the vocal tract, producing a non-periodic, noise-like signal. The vocal tract is usually modelled as an all-pole filter and the lip radiation as an FIR first-order differentiator. Since the source-filter model is a linear model, the lip radiation effect can be combined with the glottal flow resulting in the GFD. For a single pitch period, the glottal flow and the GFD are called the *glottal pulse* and the *glottal pulse derivative* waveform, respectively.

[1] http://cas.et.tudelft.nl/~andreas/matlab_code/GEFBA.rar

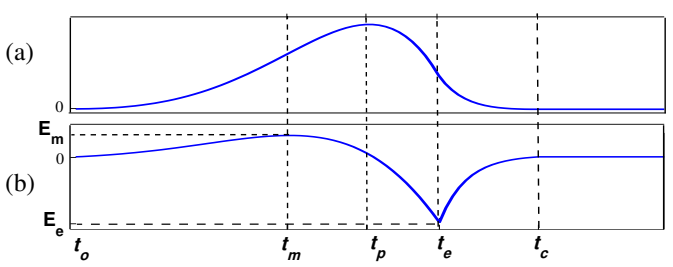


Fig. 2. The glottal parameters are illustrated for (a) the glottal pulse and (b) the glottal pulse derivative. t_o denotes the glottal opening instant (GOI), t_m is the time instant of the maximum value of the glottal pulse derivative, t_p is the fist zero-crossing instant (FZCI), t_e corresponds to the glottal closure instant (GCI), and t_c denotes the end of the return-phase.

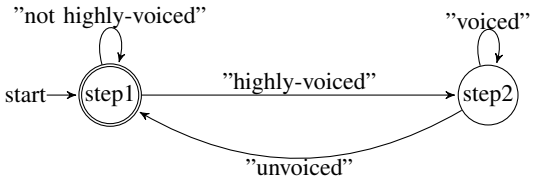


Fig. 3. Finite state machine of phase 2 of the GEFBA algorithm: step 1 searches for a highly-voiced frame and its corresponding glottal parameters. When a highly-voiced frame is found, step 2 finds the remaining glottal parameters of the voiced parts to the left and right of the highly-voiced frame until it reaches the neighbouring unvoiced segments to the left and right of the current voiced segment. Subsequently, step 1 again starts searching for a highly-voiced frame to the right of the completed voiced segment. The whole procedure is terminated when the end of the entire speech signal is reached.

The glottal pulse and its derivative can be separated into three main time-domain regions, according to the state of the vocal folds: the open-phase, the return-phase and the closed-phase, which correspond to the situation where the vocal folds are opening, closing, and remain closed, respectively. A well-known model for the coarse structure of the glottal pulse derivative is the Liljencrants-Fant (LF) model [5]. Figure 2 illustrates the glottal pulse derivative according to the LF model (lower pane) and the corresponding glottal pulse obtained by integrating the LF waveform (upper pane). The time instant of the open-phase initiation is called the *glottal opening instant (GOI)* and is denoted by t_o, while the time instant t_e at which the glottal pulse derivative reaches its minimum value, E_e, is called the *glottal closure instant (GCI)*. The time instant at which the glottal pulse derivative takes its maximum value, E_m, is denoted by t_m. The return-phase starts at the GCI and ends at t_c where the closed-phase starts. The effective duration of the return-phase is denoted by t_a and it is less than $t_c - t_e$. The *first zero-crossing instant (FZCI)* on the left of the GCI is denoted by t_p. Finally, the time interval between two successive GCIs is one pitch period long and is denoted by d_e. We will refer to the parameters t_o, E_e, t_e, t_a, t_p, t_c, E_m, t_m and d_e as the *glottal parameters*.

It should be noted that the GCI t_e is defined as the instant of significant excitation of the vocal tract [48], i.e. the instant of the negative peak of the GFD, while t_c is the instant at which the glottal flow reaches zero level. As can be seen in Figure 1, the dEGG peaks can be used as a reference for GOI and GCI extraction [49], denoting the instant of significant increase and decrease of the glottal flow, respectively [50]. The large dEGG peaks correspond to the t_e instants (see Figure 1).

GEFBA estimates all glottal parameters (in voiced segments only), except for t_c and t_a. We refer to a voiced *frame* as a fixed interval of voiced speech, while voiced *segments* refer to speech regions of various lengths. When we refer to unvoiced segments we do not necessarily mean unvoiced speech; it can also be silence. A *highly-voiced* segment/frame consists of very similar neighbouring glottal pulse derivatives. This similarity is defined in Section III-C.

The reason that GEFBA does not estimate t_c and t_a is that the VUD and GCI/GOI estimation do not depend on them but on the remaining glottal parameters. There are several ways of estimating t_c. In [8] t_c is obtained by $t_c = t_e + 1$, a value that is documented in the work of Rosenberg [3]. However, the Rosenberg model does not model the return-phase and, therefore, is less accurate than the LF model. Another more elegant method [51] proposed t_c to be the time instant at which the glottal pulse derivative returns to zero after the t_p instant. Both methods can be easily implemented in the GEFBA framework. The t_a instant can be obtained by more complex LF model fitting techniques [9].

III. The GEFBA Algorithm

The GEFBA algorithm consists of two main phases. In phase 1, described in Section III-A, a rough approximation of the source signal of the entire speech signal is obtained. In phase 2, described in Section III-E, simultaneous VUD and GCI/GOI estimation is performed in two steps. In step 1, GEFBA searches for a highly-voiced speech frame and estimates its glottal parameters. In most cases, the selected voiced frame is part of a longer continuous voiced segment and, therefore, the remaining glottal parameters to the left and right of the highly-voiced frame should be estimated. Step 2 successively "fills in the gaps" to the left and right of the highly-voiced frame. The finite state machine illustrated in Figure 3 summarizes phase 2.

The method for glottal parameter estimation of one pitch period is outlined in Section III-B. Furthermore, in both steps, the glottal parameters are successively estimated, one pitch period at a time, until an unvoiced segment is reached. Therefore, GEFBA is a pitch-period-based VUD with high resolution at the boundaries of the voiced segments. The proposed VUD algorithm consists of a set of conditions described in Section III-C. Finally, in both steps two main procedures take place: *move forward (MF)* and *move backward (MB)* explained in Section III-D. The main purpose of both functions is to find all glottal parameters left and right of an already found GCI.

A. Phase 1: Estimation of the GFD

It is well-known that the GFD can be accurately modeled as a mixed-phase signal [47], [52]. According to the frequency domain point of view of the LF model [5], [53], the open-phase corresponds to a maximum-phase component called the *glottal formant* which is modelled as two complex poles outside the unit circle and close to the real axis with a frequency of $f_p = 1/(2(t_p - t_o))$ [5]. The maximum-phase component has approximately a -12 dB/octave spectral-magnitude

roll-off. The return-phase corresponds to a minimum-phase component which can be approximated by a low-pass filter having a -6 dB/octave spectral-magnitude roll-off after a cut-off frequency $f_a = 1/(2\pi t_a)$ [5]. In total, the LF model has approximately a -18 dB/octave spectral-magnitude roll-off after f_a. It is worth noting that the Rosenberg model [3] does not have this extra -6 dB/octave roll-off since it assumes that the return-phase has zero length. Moreover, the lip radiation filter can be approximated as a high-pass filter with a $+6$ dB/octave spectral-magnitude increment. Therefore, the GFD contributes to the speech spectrum a roll-off of -12 dB/octave after f_a.

In [54] it was experimentally shown that for several types of voiced speech (i.e., modal, breathy, falsetto, vocal fry) t_p is considerably larger than t_a which means that f_a is expected to be greater than f_p. According to this assumption, we can summarize that the GFD between f_p and f_a has a -6 dB spectral-magnitude roll-off, while for frequencies greater than f_a the spectral-magnitude roll-off is -12 dB.

As explained in Section I, a smooth approximation of the coarse structure of the GFD is very convenient for estimating the GCIs/GOIs even in segments with low vocal intensity. Therefore, high frequency components should not be present in this smooth GFD approximation. We used a simple source estimation scheme based on LP inverse filtering. This provides a smooth approximation of the GFD during voiced speech. Of course there are much more accurate GFD estimation methods in the literature [8]–[11]. However, our aim here is to obtain a fast and convenient approximation of the GFD in the context of GEFBA.

A pre-emphasis filter is often used for spectral equalization before LP. In literature, a first-order pre-emphasis filter which has a $+6$ dB/octave increment is commonly used [55]–[57]. This type of filter equalizes the low frequency content before f_a. However, after f_a, a -6 dB/octave spectral-magnitude roll-off remains. Thus, LP will better estimate the lower frequency content than the higher frequency content because of the spectral matching property [55]. Instead, here we use a second-order pre-emphasis filter, $D(z) = \left(1 - \alpha z^{-1}\right)^2$, where $\alpha = 0.99$. This filter has a $+12$ dB/octave spectral-magnitude increment and, thus, better equalizes the -12 dB/octave slope of GFD after f_a than the more commonly used first-order filter [58]. Moreover, it better de-emphasizes the lower frequencies, in the neighbourhood of the glottal formant, than the first-order pre-emphasis filter. This means that the higher frequency content will be better estimated and removed during inverse filtering.

Typically, the LP order used in the literature [11], [17], [59], is equal to or slightly greater than $f_s/1000$ (where f_s is the sampling frequency) when a first-order pre-emphasis filter or no pre-emphasis is used. Two important reasons for this order selection are: a) the glottal formant can be estimated and cancelled during inverse filtering if a higher order is used, and b) closed-phase analysis methods have very small closed-phase intervals (especially for female voices) which should be larger than the LP order [9]. The second-order pre-emphasis filter can greatly reduce the energy of the glottal formant compared to the first-order pre-emphasis filter. This means that it is safe to use a higher LP order (here we use $p = f_s/1000 + 16$) without worrying about estimating the glottal formant. Therefore, improved estimation of the higher frequency content can be achieved compared to the first-order pre-emphasis filter. Of course, the GFD estimate using a second-order pre-emphasis filter sometimes contains information from the lower frequency formants and that is why an accurate GFD estimation is something that cannot be claimed in this paper. However, the low frequency formants do not have a considerable effect on the average performance of GEFBA, as is evident from the results in Section IV.

Furthermore, when the speech signal is corrupted by additive noise, and especially noise with energy in the high frequencies, the increased LP order captures a portion of this noise and, therefore, noise is cancelled out during inverse filtering. This gives increased robustness to GEFBA because it maintains the clear, smooth structure of the GFD. However, if the noise is concentrated in the very low frequencies (i.e., in the region of the glottal formant), it cannot be cancelled out. The algorithm for the GFD estimation is summarized in the next five steps:

G1: Pre-emphasize the speech signal using $D(z)$.
G2: Apply a 50% overlap frame-by-frame autocorrelation LP analysis on pre-emphasized Hann-windowed speech frames of 30 ms length, estimating the corresponding vocal tract filters.
G3: Apply inverse filtering to every speech frame with the corresponding vocal tract filter, thus obtaining a pre-emphasized GFD segment.
G4: Estimate the GFD segment via filtering the pre-emphasized GFD with $1/D(z)$.
G5: Synthesize the GFD of the entire speech signal using the overlap-add method [47].

B. Glottal Parameters Estimation for a Single Pitch Period

Having estimated the GFD, $\dot{u}[n]$, let us assume that a GCI is identified. The GEFBA algorithm moves forward or backward in order to detect the next or the previous GCI of the currently identified GCI. When a new GCI, $t_e^{(i)}$, is detected, the corresponding $E_e^{(i)}$, $t_p^{(i)}$, $t_o^{(i)}$, $t_m^{(i)}$, $E_m^{(i)}$ and $d_e^{(i)}$ are estimated using the following algorithm which is similar to the algorithms proposed in [4], [51].

P1: Select $E_e^{(i)}$ as the GFD value at $t_e^{(i)}$ (i.e., $E_e^{(i)} = \dot{u}[t_e^{(i)}]$).
P2: Select $t_p^{(i)}$ as the first zero-crossing that is found on the left of $t_e^{(i)}$.
P3: Estimate $d_e^{(i)}$ as $d_e^{(i)} = |t_e^{(i)} - t_e^{(i\pm1)}|$, where $t_e^{(i+1)}$ indicates forward movement, while $t_e^{(i-1)}$ indicates backward movement.
P4: Estimate $t_m^{(i)}$ as $t_m^{(i)} = \max\{\dot{u}[t_e^{(i)} - 0.8d_e^{(i)}, \ldots, t_p^{(i)}]\}$.
P5: Select $E_m^{(i)}$ as the GFD value at t_m (i.e., $E_m^{(i)} = \dot{u}[t_m^{(i)}]$).
P6: Estimate $t_o^{(i)}$, as follows:

a) Find the closest zero/zero-crossing, $t_{o_1}^{(i)}$ on the left of $t_m^{(i)}$.
b) Search if there is any other point, $t_{o_2}^{(i)}$, that is on the right of $t_{o_1}^{(i)}$ and left of $t_m^{(i)}$, whose amplitude value is very close and less than κE_m (where $0 \leq \kappa < 1$). If there is, then select this point, otherwise select $t_{o_1}^{(i)}$ as the estimated GOI, $t_o^{(i)}$.

The choice of a $t_o^{(i)}$ with a small positive $\dot{u}[t_o^{(i)}]$ in P6 is explained as follows. Figure 2 shows that $t_o^{(i)}$ is the first zero on the left of $t_p^{(i)}$ that satisfies the inequality $t_o^{(i)} < t_m^{(i)}$. However, using the method described in Section III-A for the rough estimation of GFD, usually, a small non-zero value appears at the onset of the open-phase (i.e., at the $t_o^{(i)}$ instant). This small non-zero value is a function of E_m to guarantee scale-invariance. Here, we use $\kappa = 0.4$, which is empirically found to be a good choice for finding $t_o^{(i)}$.

C. Voicing Detection & Candidate Selection

In voiced speech, excluding pathological voices, the structure of neighbouring glottal pulse derivatives is similar and, therefore, it can be expected that the distances between neighbouring GCIs, FZCIs, and GOIs should also be similar. All three distances should be close to the pitch period. Any difference of the three distances depends, mostly, on the estimation accuracy of the GCIs, FZCIs, and GOIs. In general, GCI/GOI algorithms [31], [32] provide more accurate estimates of the GCIs than the GOIs. Therefore, in order to obtain an accurate estimate of the true pitch period, it is better to compute the distance of the neighbouring GCIs. We also observed that, by using the methodology described earlier for the estimation of the GFD, the distance of the neighbouring FZCIs is, on average, much closer to the distance of the corresponding GCIs than the one of GOIs. Furthermore, we observed that the differences of the neighbouring distances $t_p - t_e$ are similar and the amplitudes, E_e and E_m, of the neighbouring t_e and t_m instants, respectively, have small variations except of those that are at the boundaries between voiced and unvoiced segments.

These observations can be utilized for the efficient detection of GCIs and GOIs in voiced segments *only*, since the aforementioned distances will not be similar in unvoiced segments. Thus, the boundaries of a voiced segment are defined by the first and last GCI present in the segment. We assumed that consecutive voiced segments should have distance greater than 2PP_{max} (where $\text{PP}_{\text{max}} = 1/80$ seconds is the maximum assumed pitch period), otherwise they are considered as one voiced segment. An unvoiced segment lacks any GCI.

For the sake of convenience, let us now define four time distances:

$$t_e^{(i)} - t_e^{(i-1)} = d_e^{(i)}, \quad (1)$$
$$t_p^{(i)} - t_p^{(i-1)} = d_p^{(i)}, \quad (2)$$
$$t_o^{(i)} - t_o^{(i-1)} = d_o^{(i)}, \quad (3)$$
$$t_e^{(i)} - t_p^{(i)} = d_c^{(i)} \quad (4)$$

with,

$$E_e^{(i)} = \dot{u}[t_e^{(i)}], \quad (5)$$

and

$$E_m^{(i)} = \dot{u}[t_m^{(i)}], \quad (6)$$

where (i) denotes the glottal pulse index.

The following six conditions should be satisfied in voiced segments:

C1: $\alpha_1 d_e^{(i)} < d_e^{(i\pm1)} < \alpha_2 d_e^{(i)}$
C2: $\beta_1 d_e^{(i)} < d_p^{(i\pm1)} < \beta_2 d_e^{(i)}$
C3: $\gamma_1 d_c^{(i)} < d_c^{(i\pm1)} < \gamma_2 d_c^{(i)}$
C4: $\delta_1 d_e^{(i)} < d_o^{(i\pm1)} < \delta_2 d_e^{(i)}$
C5: $\epsilon_1 E_e^{(i)} < E_e^{(i\pm1)} < \epsilon_2 E_e^{(i)}$
C6: $\zeta_1 E_m^{(i)} < E_m^{(i\pm1)} < \zeta_2 E_m^{(i)}$

where $\rho = [\alpha_1, \alpha_2, \beta_1, \beta_2, \gamma_1, \gamma_2, \delta_1, \delta_2, \epsilon_1, \epsilon_2, \zeta_1, \zeta_2]$ is a set of control parameters that define ranges of the time distances defined in (2)-(7). The right-hand-side and left-hand-side components of inequalities C1, C2, and C4 employ the d_e distance because of its high accuracy in determining the pitch period, as already mentioned. The ρ vector can take values according to three different modes: "strict", "moderate", and "relaxed". These modes are set according to the similarity of neighbouring glottal pulse derivatives. Note that in all modes, $0 < \alpha_1, \beta_1, \gamma_1, \delta_1, \epsilon_1, \zeta_1 < 1$, while $\alpha_2, \beta_2, \gamma_2, \delta_2, \epsilon_2, \zeta_2 > 1$. The closer to 1 the control parameters are, the tighter the conditions become (i.e., the higher the similarity is between neighbouring glottal pulse derivatives). Finally, these conditions are also used as glottal parameters candidate selection discussed in Section III-D. Thus, these conditions are the key element of the simultaneous VUD and GCI/GOI estimation.

If at least one of the six conditions is not satisfied then it means that the candidate set of glottal parameters is not considered proper and is discarded. If all candidate sets of glottal parameters fail the condition checking, it means that a non-highly-voiced segment (if GEFBA is at step 1) or an unvoiced segment (if GEFBA is at step 2) is reached. Inside voiced segments, it is very rare not to find a candidate set that satisfies all conditions, because the structure of the neighbouring glottal pulse derivatives are similar (i.e., satisfy the six conditions) with small variations. The control parameters of these conditions have been selected in such a way that these variations are taken into account.

D. Forward-Backward Procedure

Move forward (MF) and move backward (MB) procedures lie in the core of GEFBA, since they provide the mechanism of glottal parameter estimation. Specifically, MF and MB move approximately one pitch period at a time, forward and backward on the GFD signal, respectively, and estimate the next set of glottal parameters as described in Section III-B. MF operates in the search interval $[t_e^{(i)} + \alpha_1 d_e^{(i)}, t_e^{(i)} + \alpha_2 d_e^{(i)}]$ and looks for a GCI (if any) such that itself and all the accompanying glottal parameters satisfy C1-C6. The algorithm of MF follows.

M1: Find all zero-crossings of the search interval.
M2: Find the minimum negative peak between each pair of neighbouring zero-crossings, resulting in N GCI candidates.
M3: Find the corresponding glottal parameters of the N candidates using the algorithm of Section III-B.
M4: Remove the sets of glottal parameters that do not respect at least one of the six conditions. The remaining sets are $M \leq N$. If $M = 0$, it means that either a non-highly-voiced segment is reached (if GEFBA is at step 1) or an unvoiced segment is reached (if GEFBA is at step 2).
M5: Select from the M remaining sets the set with the minimum E_e as the next set of glottal parameters.

The procedure is repeated for MB in the search interval $[t_e^{(i)} - \alpha_2 d_e^{(i)}, t_e^{(i)} - \alpha_1 d_e^{(i)}]$. Figure 4 provides an intuitive example for MF. In Figure 4(a), M1 and M2 are represented. Note that in M1 we find only two zero-crossings in the search interval and, thus, we obtain only $N = 1$ possible GCI candidate. Figure 4(b) depicts M3, while Figure 4(c) depicts M4 and M5. Note that M4 finds that the only candidate glottal parameter set satisfies all conditions and, therefore, M5 keeps this candidate set.

E. Phase 2: Estimation of Glottal Parameters of the Entire Speech Signal

Phase 2 (depicted in Figure 5) consists of two steps: step 1, where a highly-voiced frame (belonging to a longer voiced segment) and its glottal parameters are identified, and step 2, where the voiced "gaps" (left and right of the highly-voiced frame) and their corresponding glottal parameters are identified. These two steps are now described in detail.

1) Step 1: Finding a highly-voiced frame: In this step, GEFBA searches for a highly-voiced frame. We assumed that the minimum pitch period, PP_{min}, and maximum pitch period, PP_{max}, are $1/500$ and $1/80$ seconds, respectively. GEFBA takes frames with size four times PP_{max}, with overlap of 50%. If GEFBA finds a voiced segment of at least four pitch periods within a speech frame whose sets of glottal parameters respect the set of conditions (set in "strict" and "moderate" modes), then this speech frame is considered as highly-voiced and step 2 starts, otherwise step 1 continues until a highly-voiced frame is found. Moreover, a robust first pitch period estimate is obtained in step 1 which is required for phase 2 of the algorithm.

Let us assume that step 1 reaches a certain frame. The description of step 1 follows in more detail. First, the minimum negative peak of the frame is found as a starting reference candidate GCI, denoted by t_e. Then the MB procedure starts (using in M4 only Conditions C3, C5 and C6 set in the "strict" mode) with a pre-defined long search interval since there is no previous estimate of the pitch period. In this initial case, the search interval for MB becomes $[t_e - \text{PP}_{\text{max}}, t_e - \text{PP}_{\text{min}}]$ [2]. This initial search interval is long and it is likely to contain multiple pitch periods of a voiced speech segment. Therefore, in M5, we select the set of glottal parameters that have the closest candidate GCI to the current GCI. After finding an initial pitch period, MB continues (a) using now a search interval computed with the new pitch period, (b) using all six conditions (set in the "moderate" mode) on the left until it reaches a non-highly-voiced segment or the beginning of the frame. Note that by reaching a non-highly-voiced segment it does not necessarily mean that it is unvoiced. In this case, the final voiced/unvoiced decision will be taken from step 2.

When MB is terminated, MF starts from the same reference GCI as MB but moving in the opposite direction. Note that MF is not initialized with the pitch period estimate of MB. The reason is that MF should be independent from pitch-period mismatches that may happen in MB. Finally, when MF is terminated, all glottal parameters of MB and MF are gathered together forming L total sets of glottal parameters.

[2]Correspondingly, the search interval for MF becomes $[t_e + \text{PP}_{\text{min}}, t_e + \text{PP}_{\text{max}}]$

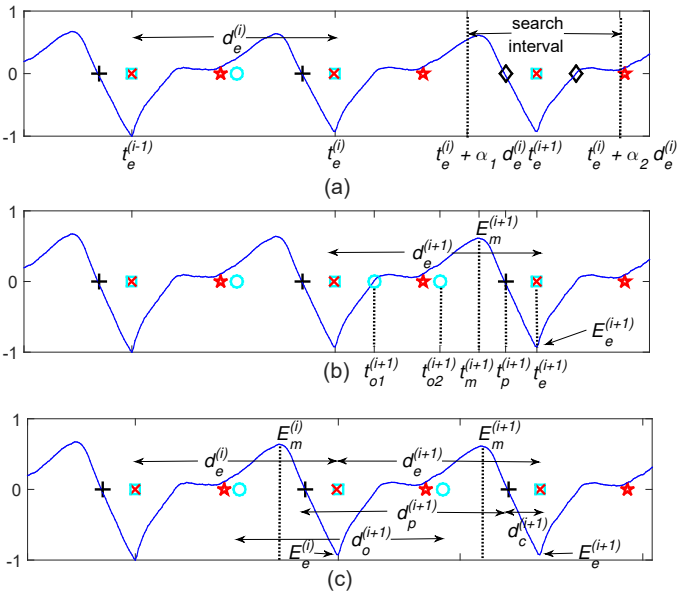


Fig. 4. Summary of move forward procedure and glottal parameter estimation for one pitch period. '+' denote the estimated first zero-crossing instants (FZCIs), 'x' and 'square' represent the true and the estimated glottal closure instants (GCIs), respectively, while 'star' and 'O' denote the true and the estimated glottal opening instants (GOIs), respectively. Finally, the 'diamond' represents found zero-crossings inside the search interval. M1: *Finding zero-crossings in the search interval*, and M2: *Finding the minimum negative peak between each pair of the estimated zero-crossings*, are described in panel (a). M3: *Apply P1-P6 to find the sets of glottal parameters* is shown in panel (b). Finally, M4: *Removing candidate sets according to C1-C6*, and M5: *Selecting the appropriate set of glottal parameters*, are depicted in panel (c).

Then, two final criteria are checked to verify if the frame is highly-voiced. The first is if $L \geq 4$, where L is the number of pitch periods in the voiced frame. It should be noted that the theoretical minimum number of consecutive pitch periods that we need in order to define periodicity is $L = 3$. We choose to use a slightly higher value in order to avoid estimating short voiced spurts. The second criterion ensures that the following inequality is satisfied

$$\frac{\min\left(d_e^{(1,2,\cdots,L)}\right)}{\max\left(d_e^{(1,2,\cdots,L)}\right)} > \lambda, \tag{7}$$

where λ is a pre-defined threshold. As previously explained, at the beginning of both MB and MF we obtain a first estimate of the pitch period which may be erroneous due to the long search interval. Therefore, the purpose of Inequality (7) is to avoid pitch-period mismatches (e.g. pitch-period halving or doubling). The pitch-period estimation mismatches occur mainly in non-highly-voiced speech. For instance, assume that MF currently analyzes the i^{th} glottal pulse derivative which is not very similar to the next, $(i+1)^{th}$, glottal pulse derivative. However it is very similar to the $(i+2)^{th}$ glottal pulse derivative. Therefore, a pitch doubling is highly probable in this case. It is empirically found that a good choice for the threshold of Inequality (7) is $\lambda = 0.6$. Theoretically, pitch halving/doubling occurs when the ratio in Inequality (7) is equal to 0.5. Thus, a slightly higher value is selected for λ in

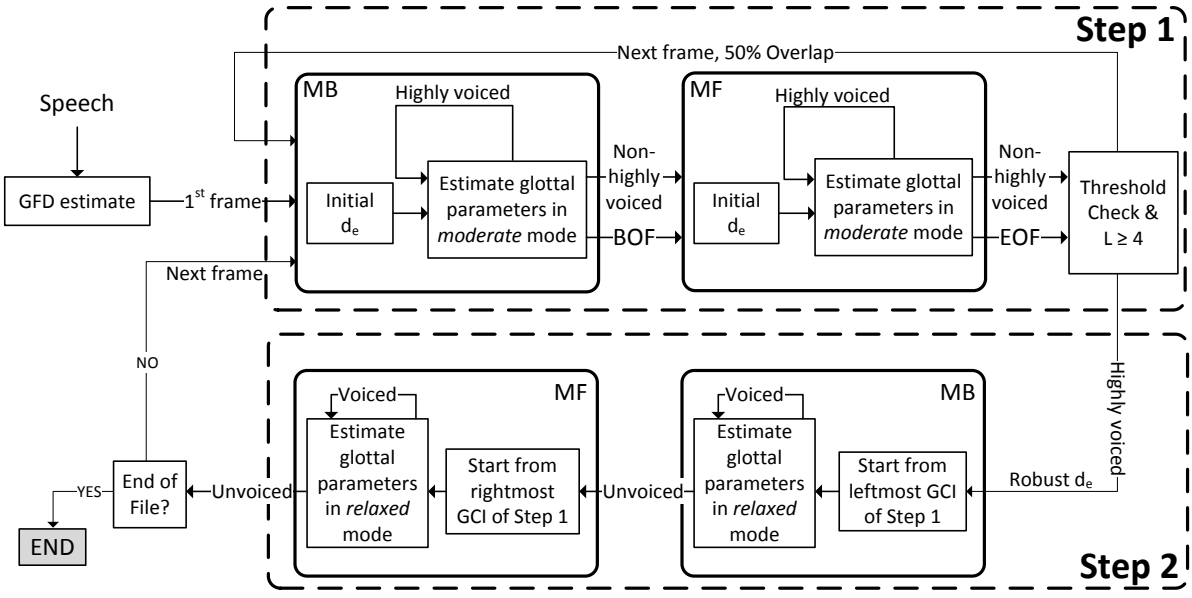


Fig. 5. Flow diagram of Phase 2: Estimation of glottal parameters of the entire speech signal by using the glottal flow derivative (GFD) of Phase 1. Phase 2 consists of two steps: a) step 1 finds a highly-voiced frame inside a larger voiced segment and estimates a robust pitch period, b) step 2 "fills the voiced-gaps" to the left and right of the highly-voiced frame. BOF is for beginning of frame and EOF is for end of frame. The pitch period estimate is denoted by d_e.

order to take into account the variations of pitch due to the quasi-periodic nature of the GFD.

MB and MF use the previous pitch period to determine the next search interval. Moreover, the two initial estimated pitch periods are prone to errors in non-higly-voiced segments as mentioned before. Therefore, if a pitch-period estimation mismatch occurs and the control parameters α_1, α_2 are set tight, then the error will continue until the procedures MB and MF terminate. In order to avoid this undesirable case, the control parameters α_1, α_2 of the "moderate" mode need to be further relaxed (even more than the "relaxed" mode). This is a necessary exception in the "strict-relaxed" paradigm, explained in Section III-C, in order to help Inequality (7) to find the pitch-period estimation mismatches.

If one of the two criteria is not satisfied the corresponding frame is considered as "non-highly-voiced" and GEFBA moves to the next frame. Otherwise it is "highly-voiced" and step 2 starts. When a highly-voiced frame and its sets of glottal parameters are found, the average pitch period is computed from the L estimated pitch periods of this frame, giving a robust pitch-period estimate that is used in step 2. Note that a new robust pitch period is estimated for each entire voiced segment.

2) Step 2: "Filling the gaps": In step 2, GEFBA moves backwards starting from the leftmost estimated GCI from step 1. At the beginning, MB uses the average pitch period from step 1 and keeps going on by replacing it with the previously estimated pitch period each time. It stops if it finds unvoiced speech (one of the five conditions is not satisfied) or if it reaches up to the first sample of the entire speech signal. Then, it collects the glottal parameters and merges them with the gathered glottal parameters of step 1. Then it starts moving forward starting from the rightmost estimated GCI obtained from step 1. Again, either it finds unvoiced speech or it keeps moving forward until it reaches the end of the entire speech signal. Then, all glottal parameters estimated in MF are collected and concatenated at the end of the other glottal parameters found so far. When we find all the glottal parameters of the entire voiced speech segment, GEFBA goes back to step 1, and the whole process starts again. This time the next frame that will be searched in step 1 starts slightly more than one minimum pitch period (i.e., $1/300$ seconds) after the rightmost previously estimated GCI of step 2. Note that in step 2 the set of conditions are set in "relaxed" mode in order to find the non-highly voiced segments as well.

IV. Evaluation

In this section, we compare the GEFBA algorithm with the "voiced-only" version (i.e. using elimination of instants that belong to unvoiced speech) of YAGA [32] and the combination of DYPSA [30] and SEDREAMS [31], [35] with two state-of-the-art VUD algorithms, RAPT [39] and SRH [41]. "Voiced-only" versions are denoted by the subscript $_V$ in the rest of the work. Although the performance of GCI/GOI algorithms without combining them with a VUD algorithm is not conclusive on which one is the best in the context of real-world applications, we also evaluate the standard versions of YAGA, DYPSA and SEDREAMS, to highlight what appears to be the bottleneck due to the combination with a VUD algorithm. These versions are named after the corresponding algorithm, with no subscript letter. The experiments are undertaken in clean and noise-contaminated speech using the parametrizations published in the corresponding papers. Moreover, the algorithms are tested using three different types of additive noise (white Gaussian noise (WGN), babble noise and car-

interior noise) taken from NOISEX92 [60] with SNR values ranging from 0 to 30 dB with a step of 5 dB.

Two databases of speech and EGG recordings were used for the evaluation. The APLAWD database [45] consists of 10 sentences repeated 5 times from 10 speakers (5 males, 5 females). The SAM database [44] consists of extended two-minute passages by four speakers (two females and two males). Reference GCIs and GOIs were extracted from the peaks of the dEGG signal via the SIGMA algorithm [49] which was experimentally shown to achieve very accurate results in APLAWD and SAM databases [49]. In order to remove any bias between estimated GCIs and reference GCIs caused by the propagation time from the glottis to the recording device, we used a constant propagation time of 0.87 ms and 0.95 ms for SAM and APLAWD, respectively. The MATLAB implementations of DYPSA and SIGMA are published by their corresponding authors in the VOICEBOX Toolbox [34], while the MATLAB implementation of SEDREAMS for GCI is published in [61]. In this implementation, we also added the GOI estimation according to [31]. Moreover, the RAPT implementation of VOICEBOX is used, while the implementation in [61] is used for SRH.

The basic component which is used in most of the evaluation measures is the glottal cycle. The glottal cycle for the i^{th} reference GCI is considered to be the interval

$$\Big[\frac{t_e^{(i-1)}+t_e^{(i)}}{2}, \frac{t_e^{(i)}+t_e^{(i+1)}}{2}\Big] \tag{8}$$

inside a voiced segment, while for the left and right boundaries we consider the intervals

$$\Big[t_e^{(i)}-\frac{t_e^{(i+1)}-t_e^{(i)}}{2}, \frac{t_e^{(i)}+t_e^{(i+1)}}{2}\Big] \tag{9}$$

and

$$\Big[\frac{t_e^{(i-1)}+t_e^{(i)}}{2}, t_e^{(i)}+\frac{t_e^{(i)}-t_e^{(i-1)}}{2}\Big], \tag{10}$$

respectively. The same intervals for the glottal cycle for the i^{th} GOI are chosen. It is reminded that the minimum distance between consecutive voiced segments is 2PP_{min}, otherwise they are considered as one. We assume that a voiced segment starts with a GCI and ends in a GCI instead of labelling frames as voiced or unvoiced, which is problematic in the boundaries. Finally, eight different evaluation metrics are used.

- Identification ratio (IDR): the percentage of glottal cycles that have exactly one GCI. IDR is a function of FAR and MR. i.e., IDR $= 100 -$ FAR $-$ MR (see below).
- False alarm ratio (FAR): the percentage of glottal cycles that have more than one GCI.
- Miss ratio (MR): the percentage of glottal cycles that have no GCI at all.
- Voiced/unvoiced detection error (VUDE): It is the proportion of samples that are erroneously classified either as voiced or unvoiced. In order to find these erroneously classified samples we apply the operation XOR to two sets A and B, where A is the set of all samples of all estimated voiced segments and B is the set of all samples of all reference voiced segments.
- Bias (in ms): the GCI bias of the error distribution after alignment.
- Std (in ms): the standard deviation of the error distribution.
- MSE (in ms): The mean square error of the error distribution which is a function of the bias and std and shows the accuracy of the algorithms.
- Relative computation time (RCT): It gives an indication of the speed of each algorithm and is computed as follows

$$\text{RCT}(\%) = 100\frac{\text{CPU}_{\text{time}}(\text{s})}{\text{Duration}_{\text{sound}}(\text{s})} \tag{11}$$

These metrics are the same as those used in [30], [32], [35] with the exception of Voiced/Unvoiced Detection Error (VUDE) which we introduce here to measure the VUD performance with high resolution. Six of the aforementioned measures (FAR, MR, IDR, Bias, Std, MSE) apply to GOI detection as well, with the only difference being the definition of a glottal cycle; t_o substitutes t_e in the intervals. The control parameter vector ρ takes the following values in the three different modes: in the "strict" mode, $\rho = [-,-,-,-,0.4,2,-,-,0.5,1,0.3,3]$, in the "moderate" mode, $\rho = [0.4,1.6,0.85,1.15,0.4,2.5,0.6,1.5,0.3,2.6,0.35,3.1]$, and in the "relaxed" mode, $\rho = [0.65,1.4,0.75,1.3,0.3,3.5,0.55,1.6,0.25,2.7,0.4,3.5]$. Note that in "strict" mode, some control parameters are not used (i.e., they are denoted by $-$) as explained in Section III-C. The selection of this parametrization is not optimal for one certain evaluation criterion but a good trade-off between them. Several sets of values are tested based on the "strict-relaxed" paradigm, and the extra relaxation of α_1 and α_2 in the "moderate" mode described in Sections III-C and III-E, respectively. The "strict-relaxed" paradigm is indeed the case for these values, except for ζ_1 which is empirically found to behave differently.

Furthermore, we excluded short voiced spurts from the evaluation by removing the reference GCIs and GOIs that are in segments with less than four pitch periods. Table I shows the performance of all algorithms in clean speech for the APLAWD and SAM databases. The entries of the table are in pairs, except for the RCT and VUDE columns. The first value of each pair is the performance for GCI estimation while the second is the performance for the GOI estimation. In RCT, we have a single value except for SEDREAMS which has a slow and a fast implementation [35]. Here, the RCT of SEDREAMS is computed by counting the time it takes for pitch estimation using the RAPT algorithm (in contrast to [35]). Figures 6, 7 and 8 show the performance of the algorithms for the three different types of additive noise using the data from both databases.

V. Discussion

In this section we discuss the performance of the compared algorithms.

A. High-Resolution VUD & Voicing Offsets

GEFBA performs high-resolution VUD because it is pitch-period-based rather than frame-based, and it is able to cap-

Database	Method	IDR	FAR	MR	VUDE	Bias	std	MSE	RCT
APLAWD	DYPSA	(96.01, 96.10)	(1.91, 1.94)	(2.08, 1.96)	37.20	(−0.07, 0.48)	(0.75, 1.00)	(0.57, 1.24)	18.5
	DYPSA$_V$	(93.88, 94.23)	(1.63, 1.64)	(4.48, 4.12)	7.76	(−0.08, 0.48)	(0.75, 1.00)	(0.58, 1.24)	58.6
	YAGA	(98.71, 98.44)	(1.10, 1.27)	(0.19, 0.28)	37.80	(−0.01, 0.76)	(0.35, 1.16)	(0.12, 1.94)	32.9
	YAGA$_V$	(86.73, 85.92)	(0.22, 0.28)	(13.05, 13.81)	9.80	(−0.02, 0.84)	(**0.30**, 1.17)	(**0.09**, 2.09)	34.3
	SEDREAMS	(97.46, 97.28)	(1.57, 1.69)	(0.97, 1.02)	39.63	(−0.06, 0.90)	(0.39, 0.99)	(0.15, 1.81)	(87.5, 58.6)
	SEDREAMS$_V$	(94.92, 94.98)	(1.34, 1.43)	(3.75, 3.59)	**7.51**	(−0.06, 0.90)	(0.40, 1.00)	(0.16, 1.83)	(87.6, 58.7)
	GEFBA	(**98.23**, **97.97**)	(**0.21**, **0.25**)	(**1.56**, **1.78**)	7.90	(**−0.01**, **−0.12**)	(0.37, **0.64**)	(0.14, **0.43**)	**15.03**
SAM	DYPSA	(94.92, 94.97)	(2.40, 2.47)	(2.67, 2.56)	51.31	(0.00, 0.69)	(0.60, 0.96)	(0.36, 1.40)	18.1
	DYPSA$_V$	(91.52, 91.90)	(1.92, 2.02)	(6.56, 6.08)	6.55	(**−0.01**, 0.68)	(0.58, **0.95**)	(0.34, 1.38)	58.9
	YAGA	(97.79, 97.38)	(1.97, 2.23)	(0.24, 0.40)	51.75	(−0.01, 0.86)	(0.36, 1.20)	(0.13, 2.20)	33.4
	YAGA$_V$	(80.06, 78.59)	(0.30, 0.40)	(19.65, 21.02)	10.29	(−0.03, 0.95)	(**0.27**, 1.23)	(**0.07**, 2.43)	92.2
	SEDREAMS	(97.14, 96.51)	(1.50, 1.83)	(1.36, 1.67)	51.82	(0.00, 1.30)	(0.39, 1.06)	(0.15, 2.83)	(75.6, 56.8)
	SEDREAMS$_V$	(93.28, 93.00)	(1.06, 1.35)	(5.65, 5.65)	**6.12**	(**−0.01**, 1.29)	(0.38, 1.09)	(0.15, 2.87)	(76.3, 56.8)
	GEFBA	(**96.72**, **96.47**)	(**0.29**, **0.34**)	(**2.99**, **3.19**)	6.29	(0.05, **0.47**)	(0.42, 0.98)	(0.18, **1.18**)	**15.6**

TABLE I

PERFORMANCE OF ALL METHODS ON CLEAN SPEECH USING THE EVALUATION CRITERIA DESCRIBED IN SECTION IV. EACH ENTRY PAIR OF NUMBERS DENOTES GCI AND GOI ESTIMATION PERFORMANCE. $_V$ DENOTES THE "VOICED-ONLY" VERSION OF THE CORRESPONDING ALGORITHM. BEST PERFORMANCES OF "VOICED-ONLY" VERSIONS ARE HIGHLIGHTED WITH BOLD.

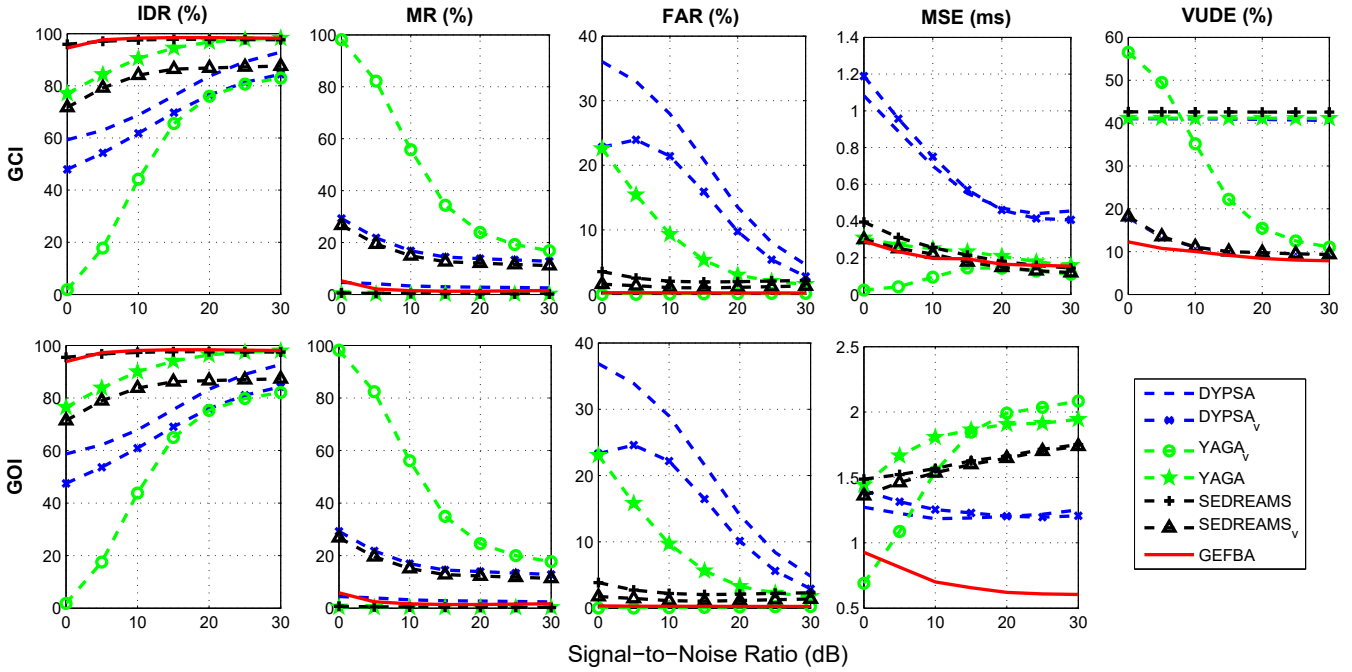


Fig. 6. VUD and GCI/GOI detection performance, of all algorithms in both (SAM, APLAWD) databases contaminated with additive white Gaussian noise, using the evaluation criteria described in Section IV.

ture the voicing offsets. Moreover, we believe that VUDE gives a high-resolution criterion about the voiced/unvoiced performance of a VUD algorithm when it is combined with a GCI/GOI algorithm. This is justified as follows. First, frame-based VUD algorithms have low resolution at voiced segment boundaries. There are two types of resolution errors that may occur. The first one is to include an unvoiced segment to a voiced one, and the second is to miss the end part of a voiced segment. The former can be resolved via the combination of a VUD with a GCI/GOI algorithm. In this combination, we can stop at the final estimated GCI and discard all the remaining unvoiced part. The latter cannot be resolved by the combination since the voiced detector labels all remaining GCIs/GOIs outside of its voiced decision interval as unvoiced. Therefore, we compare the VUDE of GEFBA with the other "voiced-only" algorithms and not RAPT or SRH themselves.

As discussed in Section I, there is a category of voicing offsets, where the speech signal remains clearly periodic for a few cycles at the end of the voiced segment while the dEGG signal is almost zero [43]. Therefore, the SIGMA algorithm does not estimate any GCIs or GOIs during this interval [49]. The main reason is that the vocal folds are "flapping in the breeze" as is stated in [43] and, therefore, they do not collide in order to produce distinguishable epochs at the

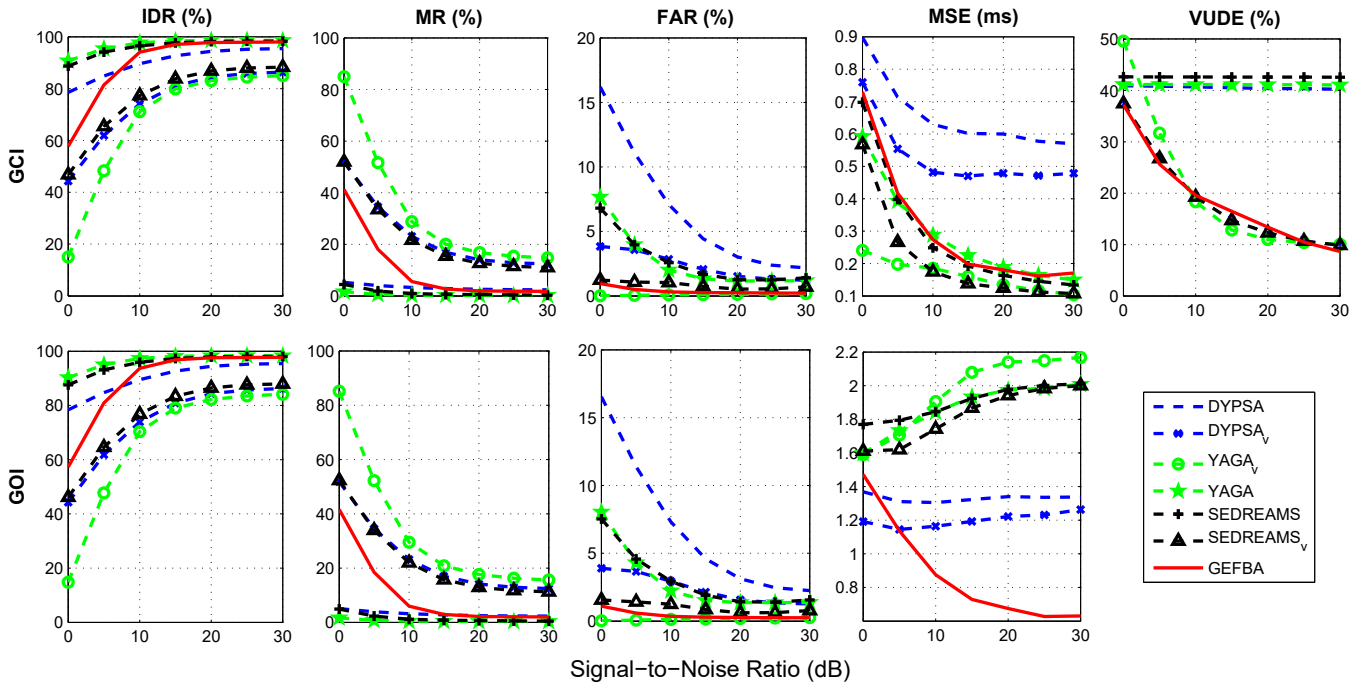


Fig. 7. VUD and GCI/GOI detection performance, of all algorithms in both (SAM, APLAWD) databases contaminated with additive babble noise, using the evaluation criteria described in Section IV.

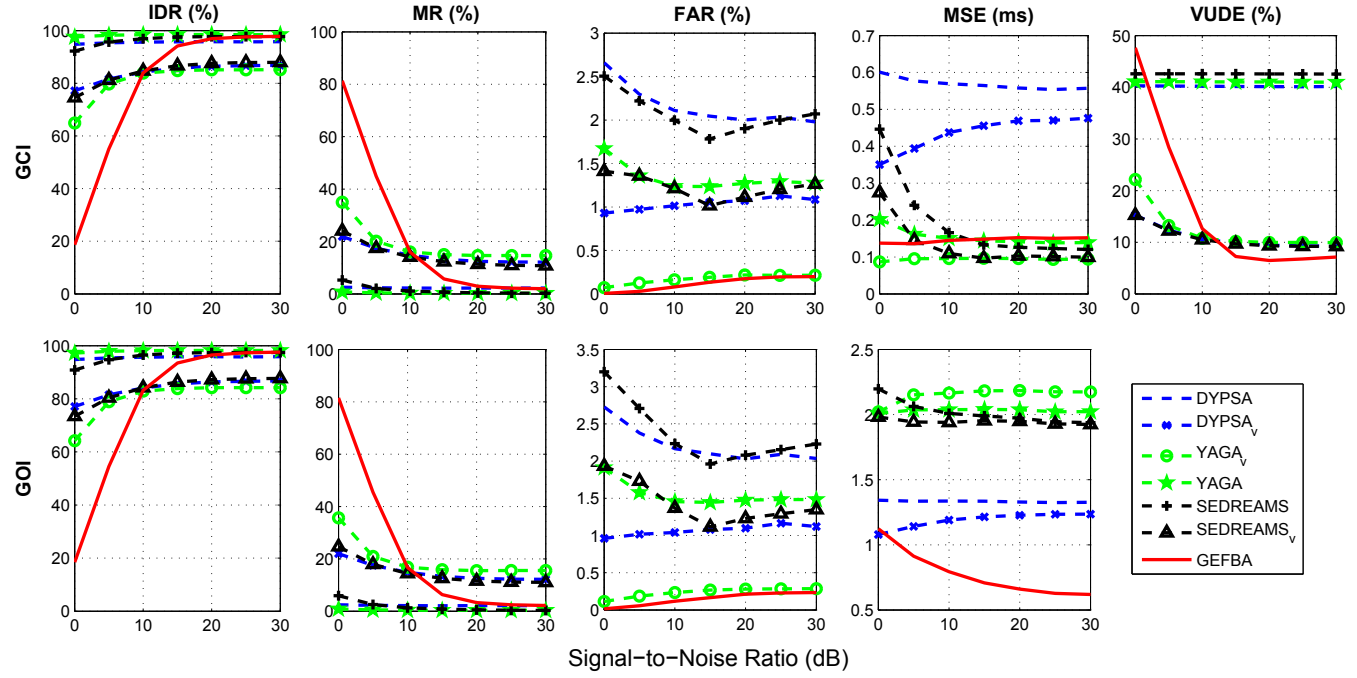


Fig. 8. VUD and GCI/GOI detection performance, of all algorithms in both (SAM, APLAWD) databases contaminated with additive car-interior noise, using the evaluation criteria described in Section IV.

dEGG waveform. This phenomenon is also called *abduction* of the vocal folds [4]. This means that the VUDE evaluation methodology based on the reference GCIs/GOIs extracted via the SIGMA algorithm is still not completely accurate. Note, however, that the systematic error (which occurs at the beginning and end of every entire voiced segment) of the frame-based-labelled VUD evaluation methods is larger than the proposed evaluation method.

We noticed that most of the VUDE of GEFBA appears in these voicing offsets and in particular in two cases: 1) the voiced-to-silence transitions and 2) the voiced-to-(unvoiced speech) transitions. In both cases, there is still some periodicity of vocal folds. In the binary VUD problem, in our opinion the first case should be considered as voiced and the second case as voiced or unvoiced. A justification of why the transition from voiced to silence should be considered as voiced is

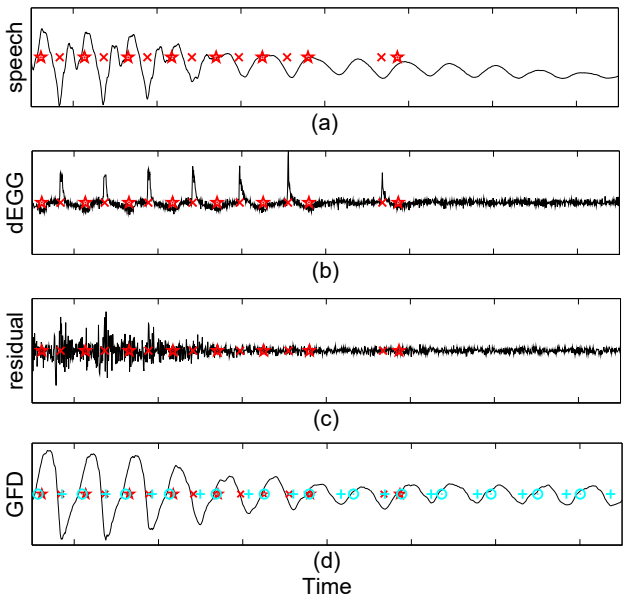


Fig. 9. *Example of voicing offsets detection using GEFBA: (a) speech segment, (b) derivative of the electroglottograph (dEGG) signal, (c) linear prediction (LP) residual, (d) GFD. Stars and 'x'-marks denote the corresponding GCIs and GOIs, respectively, as extracted from the dEGG using the SIGMA algorithm, while '+' and 'o' denote the estimated GCIs and GOIs, respectively, using the GEFBA algorithm. Clearly, the dEGG does not have sharp epochs at the voicing offset and, thus, GCIs/GOIs cannot be entirely estimated via the SIGMA algorithm.*

that in speech synthesis it is desired to model not only the main part of the voiced segment but also these voiced-to-silence transitions [4]. Therefore, we expect that the real (according to the ground-truth GCIs/GOIs extracted via a better algorithm than SIGMA) VUDE is lower than the one mentioned in Table I. A simple example that demonstrates the ability of GEFBA to capture this particular type of voicing offsets is demonstrated in 9. Clearly SIGMA cannot find this voicing offset because the dEGG does not have sharp epochs. Moreover, we observe that the last two estimated GCIs have a distance of, approximately, two pitch periods. Therefore, pitch estimation algorithms based on the dEGG [1] may miss or give erroneous pitch information at this type of voicing offsets. However, GEFBA appears not to be vulnerable in these cases.

B. Complexity

It is evident from Table I that GEFBA is much less complex than all "voiced-only" algorithms and even than the standard versions of the GCI/GOI estimation algorithms. There are three reasons for this: 1) the simple LP scheme of Phase 1, 2) the simple forward-backward movement of GEFBA in Phase 2 using simple time-domain criteria, 3) the one-step joint GCI/GOI estimation and VUD of GEFBA. The same three properties obviously hold for the noisy scenario as well. On the contrary, the higher complexity of the "voiced-only" algorithms is mainly because they perform GCI/GOI estimation and VUD in two consecutive independent steps. Moreover, DYPSA and YAGA perform N-best dynamic programming which is much more complex than the simple time-domain criteria used in GEFBA.

C. Clean Speech

Table I shows that in clean speech GEFBA outperforms the state-of-the-art in most evaluation criteria among all "voiced-only" algorithms in both databases. In the correctly identified voiced segments, it is important to have high IDR. We observe that in both databases GEFBA achieves the highest IDR among all voiced combinations in clean speech. The highest IDR occurs due to the simultaneous lowest FAR and MR. GEFBA achieves the next lower VUDE after SEDREAMS$_V$.

The simplicity of GEFBA comes with a slightly less accurate estimation of GCIs than YAGA$_V$ which performs dynamic programming. As we can see the MSE difference of GEFBA and YAGA$_V$ for clean speech is no more than 0.11 ms. Note also that YAGA$_V$ is more accurate in GCIs than YAGA. The reason is that YAGA$_V$ discards estimated GCIs from voiced segments with low vocal intensity which do not have sharp epochs and are prone to low GCI estimation accuracy. The GOI-MSE of GEFBA is remarkably better than all the other methods. The accuracy is determined from two factors: the bias and the standard deviation. The bias of the GCIs for all methods is very low because of the alignment performed. However, we can compare the bias of GOIs and GEFBA appears to have the lowest GOI bias.

D. Noise-Contaminated Speech

As discussed in Section III-A, GEFBA is based on a smooth estimated GFD removing from it any high frequency content. This is convenient for WGN and babble noise which have a portion of energy in the high frequencies. On the other hand, for car-interior noise most of the noise energy is in the low frequencies (i.e., close to f_p) and GEFBA cannot remove it from the GFD. Therefore, for this last type of noises we expect the worse performance from GEFBA.

In WGN and babble noise scenarios GEFBA achieves a much greater IDR compared to the other "voiced-only" algorithms for both low and high SNRs. In the car-interior noise scenario its performance deteriorates, as expected, for low SNR values, however for SNR > 10 dB it still outperforms the competing methods in most evaluation criteria.

In WGN, GEFBA achieves the best VUDE in all SNR values. As for the car-interior noise, GEFBA outperforms the other methods, in VUDE, only for SNR ≥ 15 dB. In babble noise scenario, GEFBA's VUDE is slightly worse than the other algorithms. This is because GEFBA sometimes captures GCIs/GOIs belonging to the noise source (babble speech) during unvoiced segments of the desired speech source.

GEFBA has a remarkably better GOI accuracy over all noise types. Furthermore, YAGA$_V$ appears to have the lowest GCI-MSE. However, for all noise types, YAGA$_V$ has an IDR that does not exceed 85% for all SNR values. The MSE difference for GCI between GEFBA and SEDREAMS$_V$ in noise-contaminated speech is less than 0.15 ms.

VI. Conclusions

In this paper we proposed the GEFBA algorithm for simultaneous voiced/unvoiced detection and estimation of the glottal closure and opening instants. Unlike other GCI/GOI

estimation algorithms, GEFBA estimates GCIs/GOIs only in voiced speech by exploiting the structure of the glottal flow derivative using simple time-domain criteria. GEFBA is also a voiced/unvoiced detector with high resolution at the boundaries of the voiced segments, since a) GFD is capable of identifying clearly the voicing offsets and b) it is pitch-period-based rather than frame-based. Common evaluation methodologies of GCI/GOI estimation algorithms do not account for possible bottlenecks in performance if combined with a VUD algorithm. Therefore, in the present paper we compared GEFBA with well-known state-of-the-art combinations of VUD and GCI/GOI algorithms. GEFBA is shown to outperform state-of-the-art combinations especially in terms of speed, GOI estimation accuracy, and identification ratio in clean speech. In additive noise scenarios, GEFBA outperforms the state-of-the art combinations in most of the evaluation criteria when the SNR is above 10 dB. A potential short-coming of GEFBA is the relative large number of control parameters, some of which may require optimization on specific type of voices. To this end, our future work includes the optimization of GEFBA's parameters, its increase in robustness for low SNRs, and its applications in speech analysis problems.

ACKNOWLEDGMENT

The authors would like to thank Dr. M. Thomas for providing the implementation of YAGA algorithm in MATLAB, and the reviewers for their helpful comments and suggestions. Finally, the authors would like to thank Mr. T. Sherson for proof-reading this paper.

REFERENCES

[1] W. Hess and H. Indefrey, "Accurate time-domain pitch determination of speech signals by means of a laryngograph," *Speech Communication*, vol. 6, no. 1, pp. 55–68, Mar. 1987.

[2] T. Ananthapadmanabha and B. Yegnanarayana, "Epoch extraction of voiced speech," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-23, no. 6, pp. 562–570, Dec. 1975.

[3] A. E. Rosenberg, "Effect of glottal pulse shape on the quality of natural vowels," *J. Acoust. Soc. Am.*, vol. 49, no. 2B, pp. 583–590, 1971.

[4] T. Ananthapadmanabha, "Acoustic analysis of voice source dynamics," *STL-QPSR*, vol. 25, no. 2-3, pp. 1–24, 1984.

[5] G. Fant, J. Liljencrants, and Q. Lin, "A four-parameter model of glottal flow," *STL-QPSR*, vol. 26, no. 4, pp. 1–13, 1985.

[6] M. Thomas, J. Gudnason, and P. Naylor, "Data-driven voice source waveform modelling," in *Proc. IEEE Intl. Conf. on Acoust., Speech, Signal Process. (ICASSP)*, Apr. 2009, pp. 3965–3968.

[7] N. D. Gaubitch and P. A. Naylor, "Spatiotemporal averaging method for enhancement of reverberant speech," in *15th International Conference on Digital Signal Processing*, July 2007, pp. 607–610.

[8] D. Y. Wong, J. D. Markel, and A. H. Gray, "Least squares glottal inverse filtering from the acoustic speech waveform," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 27, no. 4, pp. 350–355, Aug. 1979.

[9] M. Plumpe, T. Quatieri, and D. Reynolds, "Modeling of the glottal flow derivative waveform with application to speaker identification," *IEEE Trans. Speech, Audio Process.*, vol. 7, no. 5, pp. 569–586, Sep. 1999.

[10] P. Alku, "Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering," *Speech Communication*, vol. 11, no. 2-3, pp. 109–118, June 1992.

[11] P. Alku, C. Magi, S. Yrttiaho, T. Bäckström, and B. Story, "Closed phase covariance analysis based on constrained linear prediction for glottal inverse filtering," *J. Acoust. Soc. Am.*, vol. 125, no. 5, pp. 3289–3305, 2009.

[12] R. E. Slyh, E. G. Hansen, and T. R. Anderson, "Glottal modeling and closed-phase analysis for speaker recognition," in *In Proc. Speaker Odyssey: the Speaker Recognition Workshop (Odyssey 2004)*, 2004, pp. 315–322.

[13] J. Gudnason and M. Brookes, "Voice source cepstrum coefficients for speaker identification," in *Proc. IEEE Intl. Conf. on Acoust., Speech, Signal Process. (ICASSP)*, Mar. 2008, pp. 4821–4824.

[14] N. D. Gaubitch, P. A. Naylor, and D. B. Ward, "Multi-microphone speech dereverberation using spatio-temporal averaging," in *Proc. European Signal Processing Conf. (EUSIPCO)*, 2004, pp. 809–812.

[15] J. P. Cabral, S. Renals, J. Yamagishi, and K. Richmond, "HMM-based speech synthesizer using the LF-model of the glottal source," in *Proc. IEEE Intl. Conf. on Acoust., Speech, Signal Process. (ICASSP)*, May 2011, pp. 4704–4707.

[16] T. Drugman, A. Moinet, T. Dutoit, and G. Wilfart, "Using a pitch-synchronous residual codebook for hybrid HMM/frame selection speech synthesis," in *Proc. IEEE Intl. Conf. on Acoust., Speech, Signal Process. (ICASSP)*, Apr. 2009, pp. 3793–3796.

[17] P. Hedelin, "High quality glottal LPC-vocoding," in *Proc. IEEE Intl. Conf. on Acoust., Speech, Signal Process. (ICASSP)*, vol. 11, Apr. 1986, pp. 465–468.

[18] Y. Agiomyrgiannakis and O. Rosec, "ARX-LF-based source-filter methods for voice modification and transformation," in *Proc. IEEE Intl. Conf. on Acoust., Speech, Signal Process. (ICASSP)*, Apr. 2009, pp. 3589–3592.

[19] E. Moulines and F. Charpentier, "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones," *Speech Communication*, vol. 9, no. 5-6, pp. 453–467, Dec. 1990.

[20] D. Rentzos, S. Vaseghi, E. Turajlic, Q. Yan, and C.-H. Ho, "Transformation of speaker characteristics for voice conversion," in *IEEE Workshop on Automatic Speech Recognition and Understanding*, 2003, pp. 706–711.

[21] H. W. Strube, "Determination of the instant of glottal closure from the speech wave," *J. Acoust. Soc. Am.*, vol. 56, no. 5, pp. 1625–1629, 1974.

[22] C. Ma, Y. Kamp, and L. F. Willems, "A Frobenius norm approach to glottal closure detection from the speech signal," *IEEE Trans. Speech, Audio Process.*, vol. 2, no. 2, pp. 258–265, Apr. 1994.

[23] R. Smits and B. Yegnanarayana, "Determination of instants of significant excitation in speech using group delay function," *IEEE Trans. Speech, Audio Process.*, vol. 3, no. 9, pp. 325–333, Sep. 1995.

[24] P. S. Murthy and B. Yegnanarayana, "Robustness of group-delay-based method for extraction of significant instants of excitation from speech signals," *IEEE Trans. Speech, Audio Process.*, vol. 7, no. 6, pp. 609–619, Nov. 1999.

[25] A. Kounoudes, P. A. Naylor, and M. Brookes, "The DYPSA algorithm for estimation of glottal closure instants in voiced speech." in *Proc. IEEE Intl. Conf. on Acoust., Speech, Signal Process. (ICASSP)*, vol. 1, May 2002, pp. 349–352.

[26] A. P. Prathosh, T. V. Ananthapadmanabha, and A. G. Ramakrishnan, "Epoch extraction based on integrated linear prediction residual using plosion index," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 12, pp. 2471–2480, Dec. 2013.

[27] V. Khanagha, K. Daoudi, and H. Yahia, "Detection of glottal closure instants based on the microcanonical multiscale formalism," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 12, pp. 1941–1950, Dec. 2014.

[28] V. Tuan and C. d'Alessandro, "Robust glottal closure detection using the wavelet transform," in *Proc. European Conf. on Speech Communication and Technology*, Sep. 1999, pp. 2805–2808.

[29] T. Ananthapadmanabha and B. Yegnanarayana, "Epoch extraction from linear prediction residual for identification of closed glottis interval," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-27, no. 4, pp. 309–319, Aug. 1979.

[30] P. A. Naylor, A. Kounoudes, J. Gudnason, and M. Brookes, "Estimation of glottal closure instants in voiced speech using the DYPSA algorithm," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 1, pp. 34–43, Jan. 2007.

[31] T. Drugman and T. Dutoit, "Glottal closure and opening instant detection from speech signals," in *Proc. Interspeech Conf.*, Sept. 2009.

[32] M. R. P. Thomas, J. Gudnason, and P. A. Naylor, "Estimation of glottal closing and opening instants in voiced speech using the YAGA algorithm," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 1, pp. 82–91, Jan. 2012.

[33] A. Bouzid and N. Ellouze, "Glottal opening instant detection from speech signals," in *Proc. of the 12th European Signal Processing Conference*, 2004.

[34] M. Brookes, "Voicebox: speech processing toolbox for matlab," 2007. [Online]. Available: http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html

[35] T. Drugman, M. Thomas, J. Gudnason, P. Naylor, and T. Dutoit, "Detection of glottal closure instants from speech signals: A quantitative

review," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 3, pp. 994–1006, March 2012.

[36] B. S. Atal and L. R. Rabiner, "A pattern recognition approach to voiced-unvoiced-silence classification with applications to speech recognition," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-24, no. 3, pp. 201–212, June 1976.

[37] L. J. Siegel, "A procedure for using pattern classification techniques to obtain a voiced/unvoiced classifier," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-27, no. 1, pp. 83–89, Feb. 1979.

[38] D. G. Childers, M. Hahn, and J. N. Larar, "Silent and voiced/unvoiced/mixed excitation (four-way) classification of speech," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 37, no. 11, pp. 1771–1774, Nov. 1989.

[39] D. Talkin, "A robust algorithm for pitch tracking (RAPT)," *Speech Coding and Synthesis*, pp. 495–518, 1995.

[40] S. Ahmadi and A. S. Spanias, "Cepstrum-based pitch detection using a new statistical V/UV classification algorithm," *IEEE Trans. Speech, Audio Process.*, vol. 7, no. 3, pp. 333–338, May 1999.

[41] T. Drugman and A. Alwan, "Joint robust voicing detection and pitch estimation based on residual harmonics," in *Proc. Interspeech Conf.*, Aug. 2011, pp. 1973–1976.

[42] S. Gonzalez and M. Brookes, "PEFAC - A Pitch Estimation Algorithm Robust to High Levels of Noise," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 2, pp. 518–530, Feb. 2014.

[43] D. M. Howard and G. Lindsey, "Conditioned variability in voicing offsets," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 36, no. 3, pp. 406–407, Mar. 1988.

[44] D. Chan, A. Fourcin, D. Gibbon, B. Granstrom, M. Huckvale, G. Kokkinakis, K. Kvale, L. Lamel, B. Lindberg, A. Moreno, J. Mouropoulos, F. Senia, I. Trancoso, C. Veld, and J. Zeiliger, "EUROMa spoken language resource for the EU," in *Proc. European Conf. on Speech Communication and Technology*, 1995, pp. 867–870.

[45] G. Lindsey, A. Breen, and S. Nevard, "SPAR's archivable actual-word databases," Univ. College London, London, UK, Tech. Rep., 1987.

[46] G. Fant, *Acoustic theory of speech production: with calculations based on X-ray studies of Russian articulations*. The Hague, The Netherlands: Mounton, 1970.

[47] T. F. Quatieri, *Discrete-Time Speech Signal Processing: Principles and Practice*. Upper Saddle River, NJ: Prentice Hall, 2002.

[48] T. Drugman, P. Alku, A. Alwan, and B. Yegnanarayana, "Glottal source processing: From analysis to applications," *Computer Speech & Language*, vol. 28, no. 5, pp. 1117–1138, 2014.

[49] M. R. P. Thomas and P. A. Naylor, "The SIGMA algorithm: A glottal activity detector for electroglottographic signals," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 17, no. 8, pp. 1557–1566, 2009.

[50] D. Childers, D. Hicks, G. Moore, L. Eskenazi, and A. Lalwani, "Electroglottography and vocal fold physiology," *J. of Speech and Hearing Research*, vol. 33, pp. 245–254, 1990.

[51] P. Alku and E. Vilkman, "Effects of bandwidth on glottal airflow waveforms estimated by inverse filtering," *J. Acoust. Soc. Am.*, vol. 98, no. 2, pp. 763–767, Aug. 1995.

[52] B. Doval, C. d'Alessandro, and H. Nathalie, "The spectrum of glottal flow models," *Acta Acoustica United with Acustica*, vol. 92, no. 6, pp. 1026–1046, Dec. 2006.

[53] G. Fant, "The LF-model revisited. transformations and frequency domain analysis," *STL-QPSR*, vol. 36, no. 2-3, pp. 119–156, 1995.

[54] D. G. Childers and C. K. Lee, "Vocal quality factors: Analysis, synthesis and perception," *J. Acoust. Soc. Am.*, vol. 90, no. 5, pp. 2394–2410, 1991.

[55] J. Makhoul, "Linear prediction: A tutorial review," *Proceedings of the IEEE*, vol. 63, no. 4, pp. 561–580, 1975.

[56] A. H. Gray, JR. and J. D. Markel, "A spectral-flatness measure for studying the autocorrelation method of linear prediction of speech analysis," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 22, no. 3, pp. 207–216, 1974.

[57] G. Kafentzis, "On the inverse filtering of speech," Master thesis, University of Crete, Dept. of Computer Science, 2010.

[58] A. I. Koutrouvelis, "Speech production modelling and analysis," Master thesis, Delft University of Technology, 2014.

[59] J. D. Markel and A. H. Gray, *Linear Prediction of Speech*. New York: Springer, 1982.

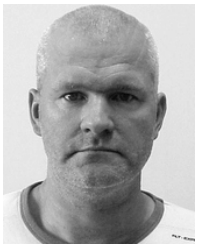
[60] A. Varga and H. J. M. Steeneken, "Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Commun.*, vol. 12, no. 3, pp. 247–251, Jul. 1993.

[61] G. Degottex, J. Kane, T. Drugman, T. Raitio, and S. Scherer, "COVAREP: A collaborative voice analysis repository for speech technologies," in *Proc. IEEE Intl. Conf. on Acoust., Speech, Signal Process. (ICASSP)*, May 2014, pp. 960–964.

Andreas I. Koutrouvelis received the B.Sc. degree in computer science from the University of Crete, Greece, in 2011 and the M.Sc. degree in Electrical Engineering from Delft University of Technology (TU-Delft), the Netherlands, in 2014. From February 2012 to July 2012, he was a research intern at Philips Research, Eindhoven, the Netherlands and from October 2014 to December 2014 he was researcher in the Circuits and Systems Group (CAS) in TU-Delft. Since, January 2015 he is pursuing the Ph.D. degree in TU-Delft (CAS). His research interests include speech analysis and multi-channel speech enhancement.

George P. Kafentzis received the B.Sc. (2008) and the M.Sc. (2010), degrees from the Computer Science Department, University of Crete, and the Ph.D. (2014) degree in Computer Science from the University of Crete and in Signal Processing and Telecommunications from the University of Rennes 1. From 2008 to 2010, he was a graduate research assistant at the Institute of Computer Science, FO.R.T.H., Heraklion, Crete, and from 2011 to 2013, he was a researcher at Orange Labs, Lannion, France. He is currently with the Department of Computer Science, University of Crete, as an adjunct lecturer. His research interests include sinusoidal modeling and modifications, inverse filtering, statistical speech synthesis, and statistical signal processing. He is an IEEE and ISCA member.

Nikolay D. Gaubitch received an MEng in Computer Engineering (1st class Honours) from Queen Mary, University of London, in 2002 and a PhD in Acoustic Signal Processing from Imperial College London in 2007. Between 2007 and 2012 he was a research associate at Imperial College London where he worked at the Centre for Law Enforcement Audio Research (CLEAR). From 2012 he has been a postdoctoral researcher with the Signal and Information Processing Laboratory (SIPLab) at Delft University of Technology where he worked on ad-hoc microphone arrays for speech enhancement in collaboration with Google, who also funded the research. His research interests span various topics in single and multi-channel speech and audio signal processing including, dereverberation, blind system identification, acoustic system equalisation and speech enhancement. He currently works as an audio researcher with Pindrop Security where his work focuses on machine learning techniques for fraud detection in the telephone channel.



Richard Heusdens received the M.Sc. and Ph.D. degrees from Delft University of Technology, Delft, The Netherlands, in 1992 and 1997, respectively. Since 2002, he has been an Associate Professor in the Faculty of Electrical Engineering, Mathematics and Computer Science, Delft University of Technology. In the spring of 1992, he joined the digital signal processing group at the Philips Research Laboratories, Eindhoven, The Netherlands. He has worked on various topics in the field of signal processing, such as image/video compression and VLSI architectures for image processing algorithms. In 1997, he joined the Circuits and Systems Group of Delft University of Technology, where he was a Postdoctoral Researcher. In 2000, he moved to the Information and Communication Theory (ICT) Group, where he became an Assistant Professor responsible for the audio/speech signal processing activities within the ICT group. He held visiting positions at KTH (Royal Institute of Technology, Sweden) in 2002 and 2008 and is a part-time professor at Aalborg University. He is involved in research projects that cover subjects such as audio and acoustic signal processing, speech enhancement, and distributed signal processing for sensor networks.